SIIM CAIMI26 Conference on Artificial Intelligence in Medical Imaging
OCT 26-27 | University of Pennsylvania, Philadelphia, PA

# General-Purpose Large Language Models Outperform Domain-Specific Models for High-Resolution CT Report Feature Extraction

Darron King, Thomas Jefferson University; Ethan Tai; Valeria Santibañez, MD; Paras Lakhani, MD; Adam Flanders, MD, FSIIM; Baskaran Sundaram, MD; Steven Rothenberg, MD

**Introduction/Background**
Large language models (LLMs) are increasingly used to extract structured clinical information from free-text radiology reports. Domain-specific medical LLMs are often presumed to outperform general-purpose models for clinical NLP tasks, yet direct benchmarking on structured radiology information extraction remains limited. We evaluated general-purpose and medical LLMs for automated extraction of report-documented high-resolution computed tomography (HRCT) features relevant to fibrotic interstitial lung disease.

**Methods/Intervention**
For this HIPAA-compliant, IRB-exempt, retrospective study, 133 HRCT reports from a single institution were manually annotated by two medical students as the reference standard. Several general-purpose (n=6) and domain-specific (n=5) open-weight LLMs were evaluated using zero shot single-prompt and multi-prompt extraction strategies. Extracted outputs included report-documented pulmonary findings, interstitial lung disease patterns, radiologic stability, and differential diagnoses. Performance was assessed using complete-report exact-match accuracy and overall finding-level accuracy. Differences between general-purpose and medical models were assessed using logistic regression with clustering by report and model. Mann–Whitney U tests were also performed.

**Results/Outcome**
General-purpose LLMs significantly outperformed domain-specific medical LLMs for overall finding-level accuracy (Figure 1, OR 2.77, 95% CI 1.51–5.08, p=0.001) and complete-report exact-match accuracy (Figure 3, OR 14.15, 95% CI 3.60–55.66, p< 0.001). Gemma 4 achieved the highest overall finding-level accuracy (93.4%), followed by Qwen (92.6%) and Llama 3 (86.7%). Restricting analyses to valid structured outputs changed complete-report exact-match accuracy by a mean difference of 0.05% across models.

**Conclusion**
General-purpose LLMs demonstrated significantly higher overall finding-level extraction accuracy than domain-specific medical LLMs. Controlling for valid structured output format had minimal effect suggesting that JSON failures do not explain the observed performance differences. These findings suggest that instruction-following capability and prompting strategy may be more important than domain-specific pretraining for structured radiology report information extraction.

**Statement of Impact**
Reliable automated extraction of structured HRCT findings from free text radiology reports could enable scalable retrospective research, AI-based peer learning, and cohort identification for clinical trials. This study

challenges the assumption that domain-specific medical models outperform general purpose LLMs and helps provide practical guidance for selecting models for radiology information extraction.

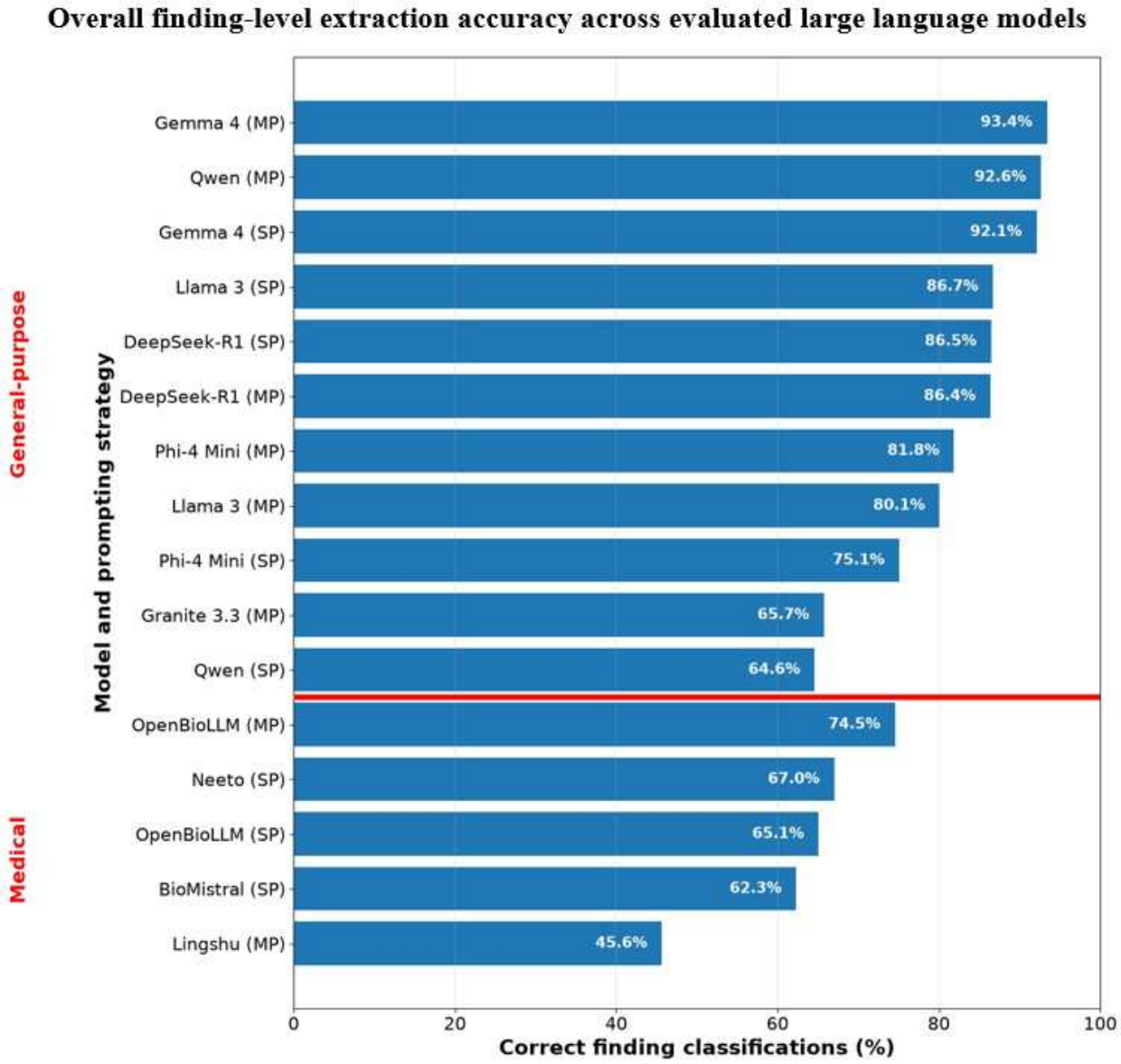


Figure 1. Overall finding-level extraction accuracy for each evaluated large language model and prompting strategy. Accuracy represents the percentage of correctly extracted report-documented pulmonary findings relative to the manually annotated reference standard. Models are grouped into general-purpose and domain-specific medical LLMs and ordered by overall finding-level extraction accuracy within each group. SP = single-prompt extraction; MP = multi-prompt extraction.

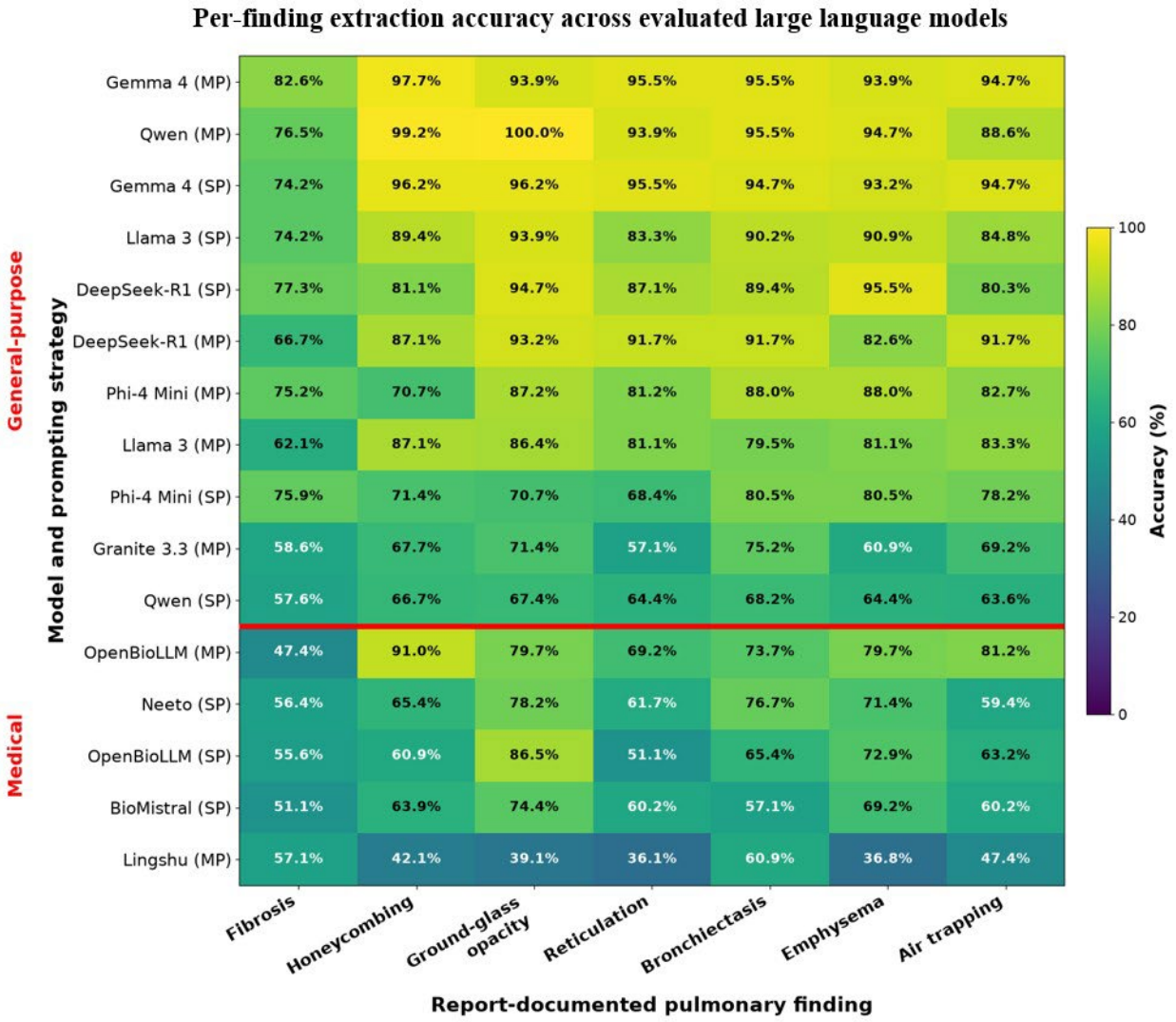


Figure 2. Extraction accuracy for each report-documented pulmonary finding by model and prompting strategy. Each cell represents the percentage of reports for which the extracted finding matched the manually annotated reference standard. Models are grouped into general-purpose and domain-specific medical LLMs and ordered by overall finding-level extraction accuracy within each group. SP = single-prompt extraction; MP = multi-prompt extraction.

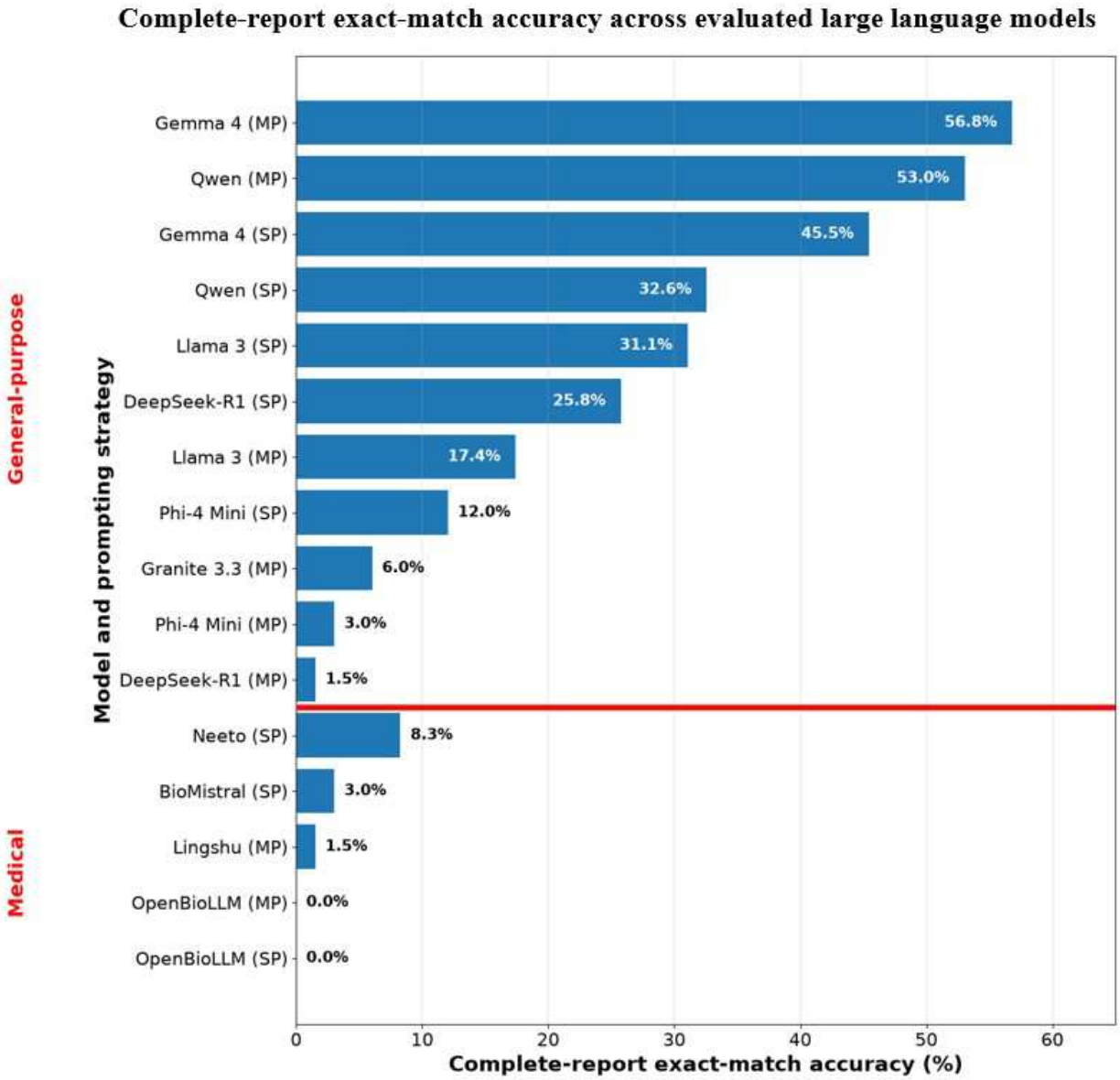


Figure 3. Complete-report exact-match accuracy for each evaluated large language model and prompting strategy. Accuracy represents the percentage of HRCT reports for which all extracted report-documented pulmonary findings, interstitial lung disease patterns, radiologic stability, and differential diagnoses exactly matched the manually annotated reference standard. Models are grouped into general-purpose and domain-specific medical LLMs and ordered by complete-report exact-match accuracy within each group. SP = single-prompt extraction; MP = multi-prompt extraction.


## Keywords

Large language models; High-resolution computed tomography (HRCT); Interstitial lung disease
Natural language processing